

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-244694

(43)Date of publication of application : 19.09.1997

(51)Int.Cl.

G10L 7/02

(21)Application number : 08-047423

(71)Applicant : NIPPON TELEGR & TELEPH CORP  
<NTT>

(22)Date of filing : 05.03.1996

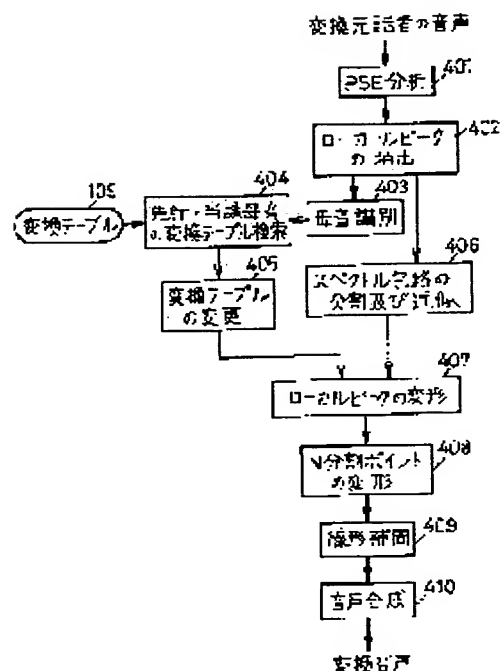
(72)Inventor : ABE MASANOBU

## (54) VOICE QUALITY CONVERTING METHOD

## (57)Abstract:

PROBLEM TO BE SOLVED: To obtain a converted speech of high quality and improve conversion efficiency.

SOLUTION: Respective spectrum envelopes are extracted from the speech of a conversion source speaker A and the speech of a conversion destination speaker B as to 5 vowels and the peak points of the respective spectrum envelopes are found; and the bands of the spectrum envelopes are divided on the basis of the frequencies at the peak points and a frequency difference and an intensity difference are found as to the division points. Each division spectrum envelope is divided into N and approximated to find an intensity differences between corresponding ones, and differences are held as a conversion table. The input speech of A is analyzed (401) to obtain a spectrum envelope, whose peak is extracted (402); and a vowel is discriminated (403) from its extracted peak. A conversion table 109 corresponding to it is taken out and used to perform deformation (407 and 408) by performing addition and subtraction among the divided spectrum envelopes (406) of the input speech and a peak point and a division point, and this deformed spectrum is used for speech synthesis (410).



## LEGAL STATUS

[Date of request for examination] 21.10.1999

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 3240908

[Date of registration] 19.10.2001

[Number of appeal against examiner's decision]

**THIS PAGE BLANK (USPTO)**

of rejection]

[Date of requesting appeal against examiner's  
decision of rejection]

[Date of extinction of right]

**THIS PAGE BLANK (USPTO)**

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平9-244694

(43) 公開日 平成9年(1997)9月19日

(51) Int.Cl.<sup>6</sup>

G 1 0 L 7/02

識別記号

庁内整理番号

F I

G 1 0 L 7/02

技術表示箇所

D

審査請求 未請求 請求項の数 5 O L (全 6 頁)

(21) 出願番号 特願平8-47423

(22) 出願日 平成8年(1996)3月5日

(71) 出願人 000004226

日本電信電話株式会社

東京都新宿区西新宿三丁目19番2号

(72) 発明者 阿部 匡伸

東京都新宿区西新宿三丁目19番2号 日本

電信電話株式会社内

(74) 代理人 弁理士 草野 卓

(54) 【発明の名称】 声質変換方法

(57) 【要約】

【課題】 高品質な変換音声を得ることができ、かつ変換効率をよくする。

【解決手段】 変換元話者Aの音声と変換先話者Bの音声とを各5母音について各スペクトル包絡を抽出し、その各スペクトル包絡のピーク点を求め、そのピーク点の周波数を基準として各スペクトル包絡を帯域分割し、これら分割点について周波数差と強度差を求め、よく各分割スペクトル包絡をそれぞれN分割して、それぞれ近似し、対応するものの間の強度差を求め、これら差を変換テーブルとしてもつ。Aの入力音声を分析して(401)、スペクトル包絡を得、これのピークを抽出し(402)、その抽出ピークから母音を識別し(403)。これと対する変換テーブル109を取出し、これを用いて、入力音声の分割スペクトル包絡(406)とピーク点、分割点を加減算して変形し(407、408)、この変形スペクトルを用いて音声合成する(410)。

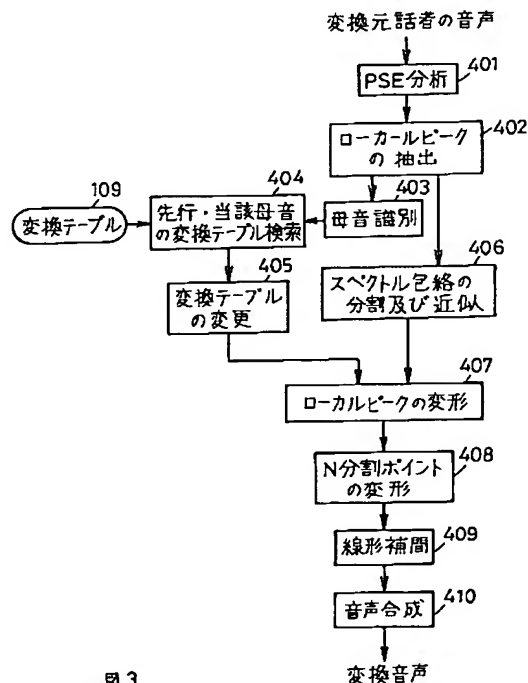


図 3

## 【特許請求の範囲】

【請求項1】 第1話者の音声を入力して、その音声を第2話者の音声へと変換する方法において、

上記第1、第2話者がそれぞれ発声した第1、第2音声から第1、第2スペクトル包絡をそれぞれ抽出するステップと、

これら抽出した第1、第2スペクトル包絡を、そのエネルギーの集中した周波数を基準にして、それぞれ複数の周波数帯域に分割するステップと、

これら分割された帯域ごとに上記第1スペクトル包絡を上記第2スペクトルへ変換するステップとを有することを特徴とする声質変換方法。

【請求項2】 上記変換するステップは上記分割された帯域ごとに上記第1、第2スペクトル包絡の変換規則を生成し、この変換規則を参照して上記第1話者の入力音声のスペクトル包絡を変形することであることを特徴とする請求項1記載の声質変換方法。

【請求項3】 上記変換するステップは音声のスペクトル空間を分割し、その分割された空間ごとに前記帯域分離されたスペクトル包絡の変換規則を用意して前記変形を行うことを特徴とする請求項2記載の声質変換方法。

【請求項4】 上記変換規則は上記第1、第2スペクトル包絡の差分であることを特徴とする請求項2又は3記載の声質変換方法。

【請求項5】 上記変換するステップにおいて、時間的に連続する変換要素を示す各上記変換規則の間を線形変換してこれら間の変換規則とする請求項2乃至4の何れかに記載の声質変換方法。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】この発明はある話者の発声した音声を入力して、その音声をあたかも特定の別人が発声したかのように変形する声質変換方法に関する。

## 【0002】

【従来の技術】例えば駅の構内放送、自動応答サービスなどでは、あらかじめ人間が発声した音声を録音しておき、その録音音声をサービス時に再生している。この場合、その1つの音声（メッセージ）内に異なる人の音声が入っていると、これを聞いた者は混乱を来すので一つの音声メッセージは同一人物によって発声されることが好ましい。一方、駅名の追加、サービスの変更などは頻繁に起こる。従って音声メッセージの追加修正は頻繁に生じる。この場合、既存の音声を発声した人物が、追加乃至変更の発声をできない場合がある。一方、音声メッセージの一部の変更又は追加のために全メッセージを発声し直すことは、多大な時間と費用を費やすことが多い。このような場合、その一部の追加、変更音声を原音声メッセージを発声した人物があたかも発声したかのように声質を変換できれば頗る便利である。このような場合に限らず、ある人が発声した音声をあたかも他の特定

の人物が発声したかのように声質を変換できれば便利な場合もある。

【0003】声質を変換する従来の方法として、スペクトル包絡からフォルマント周波数を抽出して変換する方法（例えば、文献1）と、スペクトル包絡を1つのベクトルと考え、ベクトルのマッピングによって変換する方法（例えば、文献2）がある。前者は、フォルマントの間での相関関係があるため、必ずしも希望するフォルマントが実現できるわけでは無く、高品質な変換音声を得ることは困難である。また、後者は、スペクトルの包絡の全体の歪を扱っているため、変換の効率が上がらず、高品質な音声を得るには至っていない。

【0004】（文献1）桑原、大串、“ホルマント周波数、バンド幅の独立制御と個人性判断、”電子通信学会論文誌、Vol. J69-A No. 4, pp. 509-517 (1986)

（文献2）M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, “Voice conversion through vector quantization,” J. Acoust. Soc. Jpn. (E) 11, 2, pp. 71-76 (1990)

## 【0005】

【発明が解決しようとする課題】上記のように、従来の声質変換方法は、声質の変換性能の点において、十分であるとは言えない。この発明の目的は、より精度良く声質を変換することができる声質変換方法を提供することにある。

## 【0006】

【課題を解決するための手段】この発明によれば第1、第2話者がそれぞれ発声した第1、第2音声から第1、第2スペクトル包絡をそれぞれ抽出し、これら抽出した第1、第2スペクトル包絡を、そのエネルギーの集中した周波数を基準としてそれぞれ複数の周波数帯域に分割し、これら分割された帯域ごとに第1スペクトル包絡を第2スペクトル包絡に変換する。

【0007】この変換は分割された帯域ごとに第1、第2スペクトル包絡の変換規則を生成し、この変換規則を参照して行う。この変換規則は第1、第2スペクトル包絡の差分を用いることができる。音声のスペクトル空間をクラスタリングにより複数の帯域に分割し、その各分割された帯域ごとに前記帯域分割されたスペクトル包絡の変換規則を用意して前記変換を行う。

【0008】時間的に連続する変換要素の各変換規則間を線形変換して、これら間における変換規則とする。

## 【0009】

【発明の実施の形態】次にこの発明の実施例を説明する。この実施例においてはまず変換テーブルを作成し、その変換テーブルを用いて、声質変換を行う。図1に変換テーブルの作成方法の処理手順を示す。変換元話者

A、変換先話者Bがそれぞれ発声した定常各母音をそれぞれPSE分析部101、102でPSE分析してスペクトル包絡をそれぞれ抽出する。このスペクトル包絡の抽出はPSE分析による場合に限らず、LPC分析、ケブストラム分析などスペクトル包絡を抽出できる方法であればどのようなものでもよい。

【0010】次にローカルピーク抽出部103、104で、分析部101、102でそれぞれ抽出されたスペクトル包絡のピークを見つけ、そのピークにおける周波数をローカルピーク周波数と呼ぶ。このピークの抽出は、両入力音声をLPC分析して極周波数を求め、そのバンド幅の狭い極を選択し、この周波数近傍における（例えばその前後10ポイント）PSEスペクトル包絡上のピークをそれぞれ求めて行ってもよい。また各母音ごとに有音区間におけるスペクトル包絡とローカルピーク周波数の各平均値を求め、これらを用いることができる。

【0011】スペクトル包絡分割及び近似部105、106でこれら抽出されたローカルピーク周波数をそれぞれ基準として、対応スペクトル包絡を分割する。即ち図2Aに示すようにスペクトル包絡11を、その各ローカルピーク周波数 $f_{p1}$ 、 $f_{p2}$ … $f_{p5}$ の位置で帯域分割して部分包絡 $11_1$ 、 $11_2$ 、… $11_5$ とする。これら帯域分割された部分包絡 $11_1$ 、 $11_2$ 、… $11_5$ をそれぞれN点の代表値で近似する。つまり各部分包絡をそれぞれその周波数帯域をN+1等分し、その各分割周波数における包絡値をそれぞれ代表値とする。

【0012】このようにして変換元話者Aのスペクトルの部分包絡 $11a_1$ 、 $11a_2$ 、…変換先話者Bのスペクトルの部分包絡 $11b_1$ 、 $11b_2$ 、…について、ローカルピーク間差分計算部107でローカルピーク $f_{pa1}$ 、 $f_{pa2}$ …と $f_{pb1}$ 、 $f_{pb2}$ 、…との各対応するスペクトル周波数の差 $f_{pa1} - f_{pb1}$ 、 $f_{pa2} - f_{pb2}$ 、…と、そのスペクトル強度の差とを計算して変換テーブル109に格納する。またN割ポイント間差分計算部108で、対応する部分包絡ごとの各N点の分割点ごとのスペクトル強度の差をそれぞれ計算して変換テーブル109に格納する。つまり例えば図2Bに示すようにスペクトル包絡 $11a$ 、 $11b$ の対応ローカルピーク点 $12a_3$ 、 $12b_3$ の周波数差 $\Delta f_3$ と強度差 $\Delta E_3$ を計算し、これを各ローカルピーク点について求めてテーブル109に格納し、同様に例えば部分包絡 $11a_4$ 、 $11b_4$ の各2番目の代表点 $13a_{42}$ 、 $13b_{42}$ のスペクトル強度差 $\Delta E_{42}$ を計算し、これを各対応代表点ごとに行ってテーブル109に格納する。なお各部分包絡の代表点をN点によって等分割して近似したが、2次関数、3次関数、スプライン関数などの近似によってもよい。

【0013】上述したように母音ごとに作られた変換テーブル109を用いて変換元話者Aの音声を変換先話者Bの音声に声質変換を行う手順を図3を参照して説明する。入力された変換元話者Aの音声をPSE分析部40

1でPSE分析を行いスペクトル包絡、基本周波数、有声、無声判別を求める。その求めたスペクトル包絡のローカルピークをローカルピーク抽出部402で抽出する。この抽出は図1中のローカルピーク抽出部103、104と同様の処理により行えばよい。

【0014】この抽出されたローカルピークの情報を用いて、母音識別部403で何れの母音であるか識別を行い、その識別された母音に対する変換テーブルをテーブル検索部404により変換テーブル109中から取り出す。入力音声で母音が連続する場合は、その両母音の変換テーブルをそれぞれ取り出し、変換テーブル変更部405でこれら変換テーブル間を線形変換して、両変換されるべき母音のスペクトル包絡に対する変形処理の各演算フレームにおいて用いる変換テーブルを得る。例えば入力音声でaからiと連続入力されると、図4に示すように母音aの変換テーブル21と母音iの変換テーブル22とをその入力母音a、iの入力時点 $t_1$ 、 $t_2$ の間隔をもって配し、両変換テーブル21、22の各対応するローカルピーク周波数点、また対応する分割点の値間を直線で結び、時点 $t_1$ 、 $t_2$ 間の各演算フレーム $F_1$ 、 $F_2$ 、 $F_3$ 、…における前記結んだ直線上の値を求めて、それぞれそのフレームにおける変換テーブルとする。

【0015】ローカルピーク抽出部402で抽出したローカルピークにより、その変換元話者音声のスペクトル包絡を図1中の分割近似部105、106と同様にローカルピーク周波数を基準として複数の部分包絡に分割すると共にその各分割されて得られた各部分包絡をN点で近似する。この各部分包絡の各ローカルピーク点を、ローカルピーク変形部407で変換テーブル変更部405よりの変換テーブルを参照し、対応するローカルピーク点の周波数差と強度差をそれぞれ加減算して変形する。また分割ポイント変形部408で、分割及び近似部406よりの各分割点に対し、変換テーブル変更部405よりの変換テーブルを参照して対応する点の強度を加減算して変形する。

【0016】次に線形補間部409で、先に求められたローカルピークが変形され、かつN分割点に変形された各部分包絡が線形補間されて連続した変形されたスペクトル包絡を得る。このようなことが入力音声の各母音、又は連続する母音に対して行われ、このようにして変形されたスペクトルに包絡と分析部401で得られた部分パラメータとを用いて音声合成部410で音声合成する。この音声合成法は、スペクトル包絡をゼロ位相化して、基本周波数毎に重ね合わせる方法や、スペクトル包絡から基本周波数の高周波数にわたるスペクトル強度を求め、この大きさを正弦波重量法で音声を合成する方法などで実現できる。

【0017】音声を12kHzで標本化し、16bit量子化し、PSE分析25次フレーム周期を8.0m

s、正弦波重量法で合成し、変換された音声のケプストラムと変換先音声のケプストラムとの距離を求め、変換元音声を各種変換先音声に声質変換した時の、分割点数Nに対する前記距離の変化状態をピッチが150Hzの時の実験により求めた所、図5Aに示す結果となった。この結果からNを6程度、好ましくは15程度にすれば、十分であることがわかる。他のピッチ周波数の時も同様な結果が得られた。

【0018】また発話者をよく知っている10名の者（発話者5名を含む）を被験者として、聴取音声がどの発話者のものであるか、発話者を発声者6名（発声者と被験者が同じ場合は本人を除く）選んでもらった。実際の発話者と選択話者とが一致した数を総数で割ることによって話者識別率を求め、主観評価結果とした。定常5母音音声の実験結果を図5Bに示す。ここで実際の正解率は合成音声の時の話者識別率が上限であると考えられるので、合成音声の時を100%とし、正解率を求めたものを相対比として示す。客観評価実験結果（ケプストラム距離による）を図5Cに示す。これら実験結果からこの声質変換方法が有効であることが特に基本周波数が150Hzの時には、原音声の実験と変わらない結果が得られており、この発明方法が有効であることがわかる。基本周波数が200Hzの時の結果は若干良くないが、合成音声の時に個人性に対して劣化が起っており、それに起因するものであると考えられる。特に分析条件におけるフレーム長やフレーム周期の値をすべての音声で一定にして分析、合成を行ったので、その影響がでたものと考えられる。

【0019】上述において、分割により得られた部分包絡を1つのベクトルとみなして、ベクトルのマッピングにより変換してもよい。つまり、前述した変換テーブルによる変形のみならず、分割帯域ごとにスペクトル包絡の変換規則を作り、これを参照して入力音声のスペクトル包絡を変形するようにしてもよい。更に、ベクトル量子化に用いるコードブックのように、音声のスペクトル空間を適切にクラスタリングして複数の空間に分割し、

その分割されたスペクトル空間ごとに、帯域分割されたスペクトル包絡の変換規則を用意して入力音声のスペクトル包絡を変形するようにしてもよい。変換テーブルを作成のための入力音声は定常母音のみならず、例えば発声単語から抽出したものでもよい。

#### 【0020】

【発明の効果】音声のスペクトル包絡上でエネルギーが大きい部分は、聴覚的に良く聞こえる部分であり、音声の特徴づける上で重要である。これまでの研究によれば、音声の個人性に重要であると考えられる音源の特徴も、このスペクトル包絡に反映されている。この発明では音声スペクトル包絡上でエネルギーが大きい周波数を音声の特徴量の1つとして利用し、この特徴量は、フォルマントといわれている特徴量も包含しているため、従来のフォルマント周波数のみの変換に較べて、声質変換の性能を向上させることができ、さらに、この周波数を規準として、スペクトル包絡を分割し、分割された包絡毎に変形規則を適用する。従って、スペクトル包絡を1つのベクトルとして扱って声質を変換する方式に較べ、局所的な特徴を変換することが可能となり、個人性の変換を詳細に行なうことができる。

#### 【図面の簡単な説明】

【図1】この発明に用いる変換テーブルの作成方法を示す図。

【図2】Aはスペクトル包絡の分割の方法を示す図、Bは変換テーブルを作成するためのスペクトル包絡間の差分の計算方法を示す図である。

【図3】この発明による各声質変換方法の実施例を示す図。

【図4】連続する母音の間に用いるため、両母音の変換テーブルによりテーブル変形方法を示す図。

【図5】Aは部分包絡の分割点数に対する声質変換音声及び変換先音声間のケプストラム距離との関係の実験結果を示す図、Bは主観評価実験結果を示す図、Cは客観評価実験結果を示す図である。



【図1】

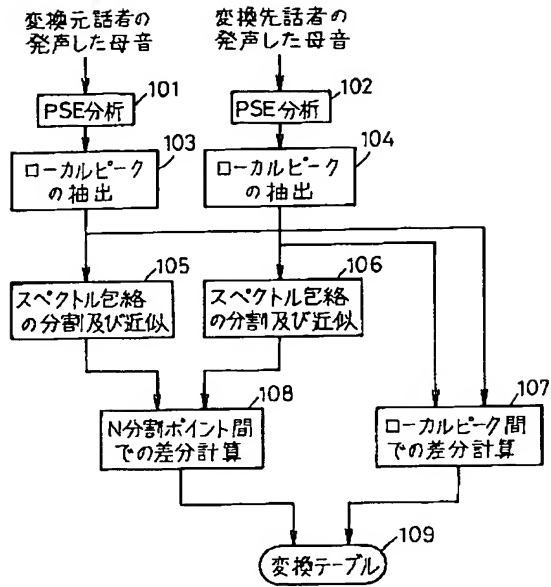


図 1

【図2】

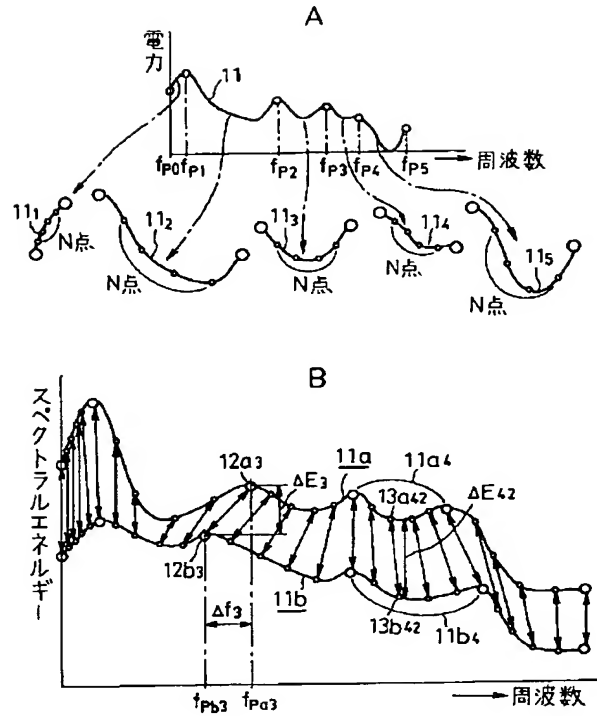


図 2

【図4】

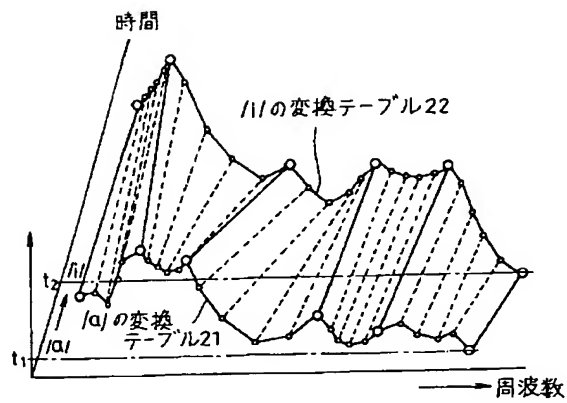


図 4

【図3】

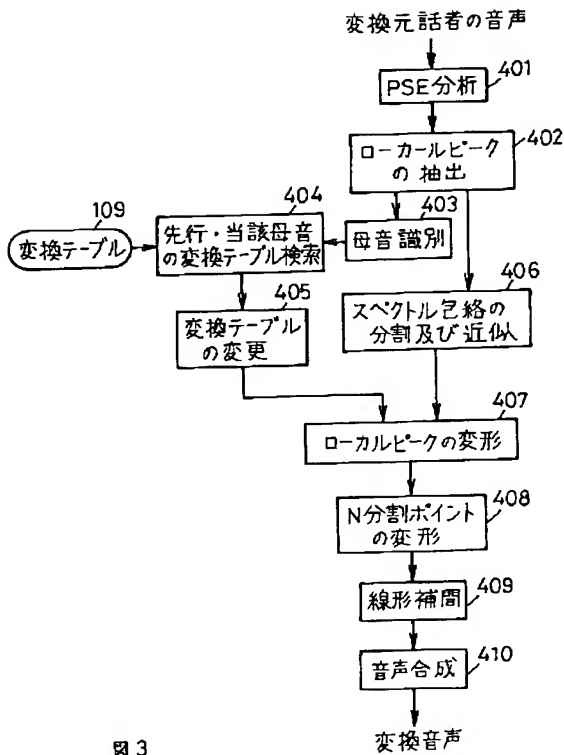


図 3

【図5】

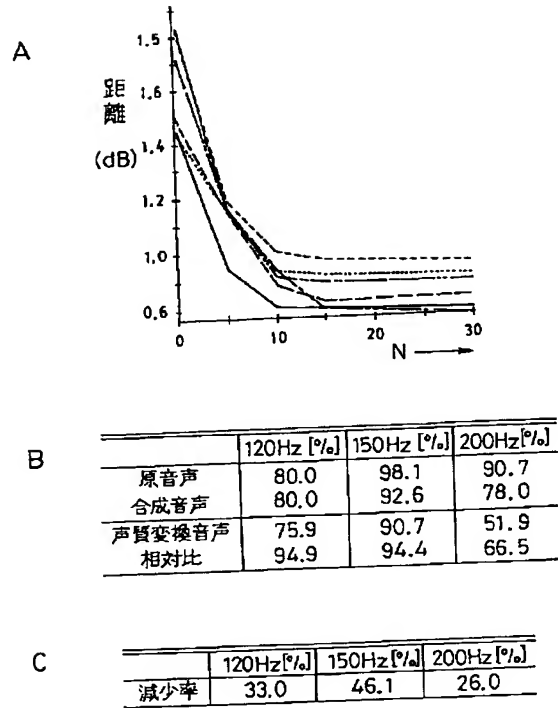


図 5